DOCUMENT RESUME

ED 127 347 TM 005 436

AUTHOR Iovett, Hubert T.
TITLE Elaboration and Application of a Theory of Criterion-Referenced Reliability.
PUB DATE [Mar 75]
NOTE 13p.; Paper presented at the Annual Meeting of the Southeastern Psychological Association (Atlanta, Georgia, March 1975)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Comparative Analysis; *Criterion Referenced Tests; Mathematical Models; Norm Referenced Tests; Statistical Analysis; *Test Reliability
IDENTIFIERS Test Theory

ABSTRACT

The reliability of a criterion referenced test was defined as a measure of the degree to which the test discriminates between an individual's level of performance and a predetermined criterion level. The variances of observed and true scores were defined as the squared deviation of the score from the criterion. Based on these definitions and the model of a score on a criterion referenced test, criterion referenced reliability was defined by analogy to norm referenced test theory as the proportion of observed variance attributable to true variance. Error variance was then defined in terms of the binomial distribution, and a calculation formula for the criterion referenced reliability coefficient was developed, which was then illustrated with hypothetical data. (Author)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED127347

ELABORATION AND APPLICATION OF A THEORY OF CRITERION-REFERENCED RELIABILITY

HUBERT T. LOVETT

Mississippi State University

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Paper presented at the meeting of the Southeastern Psychological Association, Atlanta, March 1975.

TM005436

ELABORATION AND APPLICATION OF A THEORY OF CRITERION-REFERENCED RELIABILITY

A reliability coefficient can be viewed as a measure of the discriminating power of a test. In the case of a norm-referenced test, the reliability coefficient measures the degree to which the test discriminates among individuals who differ with respect to the trait measured by the test. In the case of a criterion-referenced test, as defined by Livingston (1972), the reliability coefficient measures the degree to which the test discriminates between an individual's level of performance and some predetermined criterion level, or "cutting score."

Estimation of the reliability of norm-referenced tests has been well established in theory, and the theory has generated methodology sufficient for most educational situations (Lord and Novick, 1968). However, criterion-referenced tests have become popular only in the last few years, and development of theoretical explanations of criterion-referenced reliability has failed to keep pace with the increased use of the tests (Popham, 1972). The purposes of this paper are to elaborate on a theoretical explanation of the reliability of criterion-referenced tests and to demonstrate the applicability of the theory to practical

situations. The theoretical explanation reported here follows from that of Livingston (1972). There are, however, some important modifications with respect to the definition of total variance, degrees of freedom, and the estimation of error variance.

Background and Assumptions

In norm-referenced theory reliability is a measure of how well a test can discriminate between the performance of an individual and other members of a group. In this situation variability in performance is defined in terms of the squared deviation of an individual's score from an index of group performance, the mean. Since criterion-referenced tests are designed to discriminate between an individual's performance and a predetermined criterion level of performance, variability is defined, in this situation, in terms of the squared deviation of an individual's score from the criterion. Therefore,

$$\underline{D}^2_{x_i} = (\underline{X}_i - \underline{C})^2, \qquad (1)$$

where $\underline{D}^2_{x_i}$ is the variance of the ith observed score from the criterion, $\underline{X}_i$ is the ith observed score, and $\underline{C}$ is the criterion. It also follows that

$$\underline{D}^2_{T_i} = (\underline{T}_i - \underline{C})^2, \qquad (2)$$

where $\underline{D}^2_{T_i}$ is the variance of the ith true score from the criterion, and $\underline{T}_i$ is the ith true score.

In developing an index of reliability it is necessary to make certain assumptions relative to errors of measurement. An error of measurement is defined for criterion-referenced tests as it is for norm-referenced tests; that is,

$$\underline{e}_i = \underline{X}_i - \underline{T}_i, \tag{3}$$

where $\underline{e}_i$ is the error of measurement for the ith person. It is assumed that $\underline{e}$ is a random variable $(0, \sigma^2_e)$ and that $\underline{e}$ is independent of $\underline{T}$; that is,

$$\underline{R}_{eT} = 0, \tag{4}$$

where $\underline{R}_{eT}$ is the correlation between the errors of measurement and the true scores.

Score Model and Derivations

The model for an observed score on a criterion-referenced test is as follows:

$$\underline{X}_i = \underline{C} + (\underline{T}_i - \underline{C}) + (\underline{X}_i - \underline{T}_i). \tag{5}$$

Subtracting $\underline{C}$ gives

$$\underline{X}_i - \underline{C} = (\underline{T}_i - \underline{C}) + (\underline{X}_i - \underline{T}_i). \tag{6}$$

Squaring and Summing over $\underline{n}$ (an infinitely large number) persons yields

$$\sum_{i=1}^{n}(X_i - C)^2 = \sum_{i=1}^{n}(T_i - C)^2 + 2\sum_{i=1}^{n}(T_i - C)(X_i - T_i) +$$

$$\sum_{i=1}^{n}(X_i - T_i)^2. \qquad (7)$$

Since $R_{eT} = 0$ and $\bar{e} = 0$,

$$\sum_{i=1}^{n}(X_i - C)^2 = \sum_{i=1}^{n}(T_i - C)^2 + \sum_{i=1}^{n}(X_i - T_i)^2. \qquad (8)$$

Substituting from equations 1, 2, and 3 gives

$$\sum_{i=1}^{n} D_{X_i}^2 = \sum_{i=1}^{n} D_{T_i}^2 + \sum_{i=1}^{n} e_i^2. \qquad (9)$$

Dividing by n yields

$$E(D_X^2) = E(D_T^2) + \sigma_e^2, \qquad (10)$$

where E indicates the expected value. Criterion-referenced reliability is then defined by analogy to norm-referenced reliability, thus

$$R_{cc} = \frac{E(D_T^2)}{E(D_X^2)} = \frac{E(D_X^2) - \sigma_e^2}{E(D_X^2)} = 1 - \frac{\sigma_e^2}{E(D_X^2)}, \qquad (11)$$

where R_{cc} is the criterion-referenced reliability coefficient. It can be demonstrated that, where C is equal to the population mean, the criterion-referenced reliability equals norm-referenced reliability (Livingston, 1972). Thus, norm-referenced reliability is a special case of criterion-

referenced reliability.

Application

In applying the above definition of reliability it is necessary to find suitable estimates of $\underline{E}(\underline{D}_X^2)$ and σ_e^2. Livingston's (1972) approach to application was to partition $\underline{E}(\underline{D}_X^2)$ thus,

$$\underline{E}(\underline{D}_X^2) = \sigma_x^2 + (\underline{\mu} - \underline{C})^2, \tag{12}$$

where σ_x^2 is the population variance in the norm-referenced sense and $\underline{\mu}$ is the population mean. With a suitable estimate of the norm-referenced reliability, $\underline{E}(\underline{D}_T^2)$ can then be found thus,

$$\underline{E}(\underline{D}_T^2) = \underline{R}_{XX}\sigma_x^2 + (\underline{\mu} - \underline{C})^2, \tag{13}$$

where $\underline{R}_{XX}$ is the norm-referenced reliability of the test.

Livingston's technique requires the estimation of the test's reliability by norm-referenced methods. Using these methods can add several assumptions to the total procedure. The following approach allows estimation of $\underline{R}_{CC}$ without estimating $\underline{R}_{XX}$.

Total Variance

From equations 8, 9, and 10 it can be seen that

$$E(D_x^2) = \frac{\sum_{i=1}^{n}(X_i - C)^2}{n} . \qquad (14)$$

If a sample were used to estimate $E(D_x)$, and if Livingston's approach were used, then

$$d_x^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2 + n(\bar{X} - C)^2}{n - 1} \qquad (15)$$

where d_x^2 is the sample estimate of $E(D_x^2)$, and $\bar{X}$ is the sample estimate of μ. Division by $n - 1$ instead of n is necessitated because μ represents an independent constraint placed upon the value of s_x^2, which is the sample estimate of σ_x^2.

However, since variability in the criterion-referenced sense has been defined as $(X_i - C)^2$ and since C is not a constraint placed upon the value of d_x^2, d_x^2 can be found directly from the right side of formula 14 without loss of a degree of freedom. Therefore,

$$d_x^2 = \frac{\sum_{i=1}^{n}(X_i - C)^2}{n} . \qquad (16)$$

Error Variance

Let p_{ij} be the probability that person i will answer question j correctly on a test with k questions. If it is

assumed that $\underline{p}_{i1} = \underline{p}_{i2} = \ldots = \underline{p}_{ik} = \underline{p}_i$, then

$$\underline{T}_i = k\underline{p}_i, \tag{17}$$

and

$$\begin{aligned}\sigma^2_{e_i} &= k\underline{p}_i(1 - \underline{p}_i) \\ &= \underline{k}(\underline{p}_i - \underline{p}_i^2) \\ &= \underline{T}_i - \underline{T}_i\underline{p}_i \\ &= \frac{k\underline{T}_i - \underline{T}_i^2}{\underline{k}} \quad .\end{aligned} \tag{18}$$

Lord (1957) demonstrated that the observed score $\underline{X}_i$ can be used as an estimate of $\underline{T}_i$ in this case. There is, however, a loss of a degree of freedom because $\underline{T}_i$ is a constraint placed upon the value of $\underline{S}^2_{e_i}$, the sample estimate of $\sigma^2_{e_i}$. Therefore,

$$\underline{S}^2_{e_i} = \frac{k\underline{X}_i - \underline{X}_i^2}{\underline{k} - 1} \quad . \tag{19}$$

If formula 19 is summed and averaged over $\underline{n}$ observations, the result will be an estimate of σ^2_e:

$$\underline{S}^2_e = \frac{\underline{k}\sum_{i=1}^{n}\underline{X}_i - \sum_{i=1}^{n}\underline{X}_i^2}{\underline{n}(\underline{k} - 1)} \quad . \tag{20}$$

Estimating $\underline{R}_{cc}$

From formulas 11, 16, and 20 it follows that

$$\underline{r}_{cc} = 1 - \frac{\underline{S}_e^2}{\underline{d}_x^2} = 1 - \frac{\underline{k}\sum_{i=1}^{n}\underline{X}_i - \sum_{i=1}^{n}\underline{X}_i^2}{(\underline{k}-1)\sum_{i=1}^{n}(\underline{X}_i - \underline{C})^2}, \qquad (21)$$

Where $\underline{r}_{cc}$ is an estimate of $\underline{R}_{cc}$. Formula 21 is a relatively convenient calculation formula and will be demonstrated in the next section.

Example

Suppose a teacher defined mastery of the weekly spelling lesson as the ability to spell 90 percent of the words from the lesson correctly. There were 40 words in the lesson, and to check for mastery, the teacher asked each of her five students to spell all 40 words in the lesson. The scores made by the students were 35, 37, 39, 39, and 40. The values needed for substitution in formula 21 are

$\underline{k} = 40,$

$\underline{C} = .9(40) = 36.$

$\sum_{i=1}^{n}\underline{X}_i = 190$

$\sum_{i=1}^{n}\underline{X}_i^2 = 7236,$ and

$\sum_{i=1}^{n}(\underline{X}_i - \underline{C})^2 = 36.$

Therefore,

$$r_{cc} = 1 - \frac{40(190) - 7236}{(40 - 1)36} = 1 - \frac{364}{1404} = 1 - .26 = .74.$$

Discussion

There are two very important questions relative to the criterion-reference reliability coefficient: (1) How is the coefficient to be interpreted? And (2) does the coefficient have any practical value? The answer to the second question depends, of course, on the answer to the first; the practical value of the coefficient depends upon what it can reveal about a test.

Hambleton and Novich (1973) implied that a criterion-referenced reliability coefficient should measure the degree to which a test can discriminate between persons with true scores on one side of the criterion and persons with true scores on the other side of the criterion. This, however, would impose a kind of norm-referenced interpretation on the coefficient, and in estimating the coefficient with a sample, it would be necessary to include subjects with true scores on either side of the criterion. It is, however, possible, and indeed likely, with some of the newer approaches to instruction to have a situation in which all true scores are on one side of the criterion.

It is clear that the coefficient estimated by (21) does not indicate the degree to which the test discriminates

between persons with true scores above and below the criterion. However, it is a measure of the divergence of the true scores from the criterion. As the true scores approach the criterion, the coefficient approaches zero, and as the true scores approach maximum divergence from the criterion, the coefficient approaches one. Therefore, the coefficient can be interpreted as a measure of the degree to which observed scores which differ from the criterion in either direction are associated with true scores which also differ from the criterion.

A "cutting score" has value only if some of the people taking the test have true scores different from the cutting score. If all of the true scores equal the "cutting score" or criterion, then the divergence of an observed score from the criterion will be due to random error. The criterion-referenced reliability coefficient, therefore, indicates the validity of using a particular criterion.

References

Hambleton, R. K. & Novich, M. R. Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 1973, 10, 159-170.

Livingston, S. A. Criterion-referenced applications of classical test theory, Journal of Educational Measurement. 1972, 9, 13-26.

Lord, F. M. Do tests of the same length have the same standard error of measurement? Educational and Psychological Measurement, 1957, 17, 510-521.

Lord, F. M. & Novich, M. R. Statistical Theories of Mental Test Scores. Reading, Mass.: Addison-Wesley, 1968.

Popham, W. J. An evaluation guidebook: A set of practical guidelines for the educational evaluator. Los Angeles: The Instructional Objectives Exchange, 1972.